Semantic Annotation of the ACL Anthology Corpus for the Automatic Analysis of Scientific Literature

Kata Gábor, Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier, Thierry Charnois

Semantic Annotation of the ACL Anthology Corpus for the Automatic Analysis of Scientific Literature

Kata Gábor*, Haïfa Zargayouna*, Davide Buscaldi*, Isabelle Tellier†, Thierry Charnois*
* LIPN, CNRS (UMR 7030), Université Paris 13 Sorbonne Paris Cité
† LaTTiCe (UMR 8094), CNRS, ENS Paris, Université Sorbonne Nouvelle - Paris 3
PSL Research University, Université Sorbonne Paris Cité
(gabor,haifa.zargayouna,davide.buscaldi,thierry.charnois)@lipn.univ-paris13.fr, isabelle.tellier@univ-paris3.fr

Abstract

This paper describes the process of creating a corpus annotated for concepts and semantic relations in the scientific domain. A part of the ACL Anthology Corpus was selected for annotation, but the annotation process itself is not specific to the computational linguistics domain and could be applied to any scientific corpus. Concepts were identified and annotated fully automatically, based on a combination of terminology extraction and available ontological resources. A typology of semantic relations between concepts is also proposed. This typology, consisting of 18 domain-specific and 3 generic relations, is the result of a corpus-based investigation of the text sequences occurring between concepts in sentences. A sample of 500 abstracts from the corpus is currently being manually annotated with these semantic relations. Only explicit relations are taken into account, so that the data could serve to train or evaluate pattern-based semantic relation classification systems.

Keywords: semantic annotation, semantic relations, ACL Anthology

1. Introduction

One of the emerging trends of natural language technologies is their use for the humanities and sciences. There is a constant increase in the production of scientific papers and experts are faced with an explosion of information that makes it difficult to have an overview of the state of the art in a given domain (Larsen and von Ins, 2010). This phenomenon gave rise to efforts from the semantic web, scientometry and natural language processing communities, trying to improve access to scientific literature. Most of these works concentrate on unfolding links between papers in order to build cartographies and analyze scientific networks based on bibliographic references and topic taxonomies (Osborne and Motta, 2012; Osborne and Motta, 2014). Shotton (2010) designed an ontology of different types of citations, while Presutti et al. (2014) aim to improve access to semantic content and inter-document navigation by identifying links between documents. Other studies focus on the evolution of a scientific domain over time (Chavalarias and Cointet, 2013; Omodei et al., 2014).

As opposed to these lines of research, our analysis focuses on the semantic content of individual scientific papers instead of inter-document links and author networks. The goal of our work is to automatically build a state of the art of a scientific domain. Focusing on the abstract and the introduction, we represent the semantic content of a paper as a set of domain-specific concepts and typed relations between them. By identifying instances of concepts and domain-specific relations, we can extract the contribution of a research paper. Applying this method to a corpus of papers, we can lay the foundation to applications for visualizing the evolution of a domain, the emergence of topics over time and to extract the state of the art or make predictions over trends (Chavalarias and Cointet, 2013; Herrera et al., 2010; Skupin, 2004).

Semantic cartographies can rely on structured resources such as ontologies. Ontologies allow a fine-tuned semantic analysis, as opposed to approaches that primarily rely on unstructured resources or terminology extraction on the fly (Omodei et al., 2014; Chavalarias and Cointet, 2013; Skupin, 2004). Sateli and Witte (2015) and Presutti et al. (2014) rely on DBPedia to identify key concepts. Of particular interest for us is that systems using an ontology can also benefit from different kinds of typed relations. While statistical systems can do well on identifying key concepts, a typology of relations is much more difficult to extract from domain corpora. In the context of our research, we are interested in exploiting the relations from external ontologies, as well as populating ontologies by discovering new types of semantic relations between concepts using pattern mining techniques (Béchet et al., 2012). The semantic analysis of scientific corpora allows to add new relation types and instances to existing ontologies (Petasis et al., 2011) or thesauri (Wang et al., 2013).

As a first step towards the automated analysis of scientific corpora, the production of an annotated gold standard corpus was undertaken. We decided to concentrate on the computational linguistics domain and use part of the ACL Anthology Corpus (Radev et al., 2009) for our purposes. This article describes the (ongoing) work on annotating this corpus to create a resource for training and evaluating relation extraction systems.

Our semantic annotations are conceived in terms of entities representing domain-relevant concepts and the possible semantic relations linking them, e.g.:

<relation type="usedfor"><arg1><entity> String models </entity></arg1> are popular in <arg2><entity> statistical machine translation </entity></arg2></relation>.

Our approach is specifically focused on scientific literature, but still generic in that it can be applied to any scientific domain for which a corpus of papers is available. It is com-

posed of two consecutive steps :

- The identification of domain-specific concepts in the papers. These concepts will be linked to external ontological resources. Concepts were automatically annotated in the totality of the corpus.
- The annotation of the semantic relations that hold between these concepts (if any) - this is done manually on the gold standard corpus. These data will serve to evaluate consecutive experiments on relation extraction and classification. A sample of 500 abstracts is currently being manually annotated.

Section 2. describes how the texts of the corpus were selected and pre-processed. Section 3. presents an analysis of the annotation of concepts and the adequacy of the available resources. Section 4. explains the two-step process followed to annotate the relations: creating a typology (4.1.) and applying it (4.2.). Finally, section 5. draws the conclusions and presents directions for future work.

2. Composition of the Corpus and Pre-processing

We decided to focus on the abstract and introduction parts of scientific papers since they express essential information in a compact and often repetitive manner, which makes them an optimal source for mining sequential patterns in further experiments.

The core of the corpus is the database pre-processed by by E. Omodei (2014), which contains the abstracts and different kinds of meta-information for 13.322 papers from the ACL Anthology. We relied on the paper IDs in the database to identify and extract the introductions for the same articles from the 2009 version of the ACL Anthology Corpus (Radev et al., 2009). As a first cleaning, CRF tagger [1] was ran on the ACL Anthology Corpus to filter out files than cannot be processed: those with a significant number of OCR errors or tokenisation difficulties. We then extracted the introductions from the remaining files using the following heuristics:

- the introduction starts with a line that contains "Introduction/INTRODUCTION" or starts with "1" or "1.1" followed by a capitalized word,
- the last line of the introduction is the one followed by a line starting with "2" or "1.2" followed by a capitalized word
- if the end of the introduction cannot be identified, the first 6 lines are kept.

The introductions were paired with the corresponding abstracts : the resulting corpus contains 4.200.000 words from 11.000 papers.

People's names and bibliographic references were automatically annotated. For bibliographic references, we adapted the specific grammar available in SxPipe (Sagot and Boullier, 2008). People's names were recognized using the Stanford NER Named EntityRecognizer (Finkel et al., 2005) and its pre-trained model for three classes (among which only PERSON was used).

[1] http://crftagger.sourceforge.net

3. A First Analysis on the Annotation of Concepts

Linking entities to external ontological resources is an important aspect in the context of our work. The corpus we prepare will serve to experiment with different methods for populating ontologies by unsupervised semantic relation extraction. These methods may rely on different kinds of information: sequential patterns, distributional vectors, but also on the semantic properties of the entities, as extracted from the ontology. Therefore, entities representing concepts were annotated based on available resources.

3.1. Goals in terms of precision/annotation density

Ontologies differ with respect to their domain and to the concepts and types of relations they encode. Two aspects come into play when evaluating an ontology as a knowledge resource: the precision and recall of the resource itself (i.e. the quality and quantity of the information it contains), and the compatibility between the resource and the corpus to which it is applied. Our focus was to explore the adequacy of different available resources for concept annotation in a scientific corpus. Brewster et al. (2004) propose data-driven methodologies to evaluate the "fit" between an ontology and a domain corpus. Unlike their study, we prioritized simple coverage over structural fit. The following three criteria were considered in accordance with our purposes.

The ***coverage*** of a resource can be interpreted in two different ways. Coverage with respect to the domain vocabulary shows the proportion of the concepts of a domain that are included in the ontology: it can be interpreted as a measure of recall. However, as we do not have a "standard" domain model, we resort to the hypothesis that the ACL Anthology corpus is representative of the domain.

This brings us to the second criterion: coverage with respect to the corpus. Brewster et al. (2004) suggest using lexical keyword extraction and query expansion to extract relevant domain vocabulary. This vocabulary will be mapped to the ontology, and the ratio of words found in the ontology gives the measure of coverage for corpus vocabulary. We propose another interpretation of coverage, as ***annotation density***: the proportion of running words that are annotated by the resource. This aspect is of particular importance in the context of pattern mining. We need to assess whether relevant patterns between concepts will be well represented in the corpus. Knowing that not every pair of entity instances participate in an explicit relation, we need several recognized entities in the same sentence and within a reasonable distance to make sure that meaningful repetitions can be spotted. Annotation density will be measured as the proportion of annotated entities per 100 words (the average length of an abstract).

The third, conflicting criterion is the ***precision*** of the annotations produced by a resource. It can be measured as the proportion of relevant annotations over the total number of annotations. This aspect is correlated with the *specificity* of the resource: a good quality specialized ontology coupled with a corpus of the same domain allows to annotate the important concepts of the domain while restricting ambi-

guities or irrelevant annotations. It is important to note that besides the quality of the resource and its adequacy with the corpus, precision also depends on the accuracy of the annotation process itself.

While general ontologies such as WordNet (Fellbaum, 1998) may ensure a good annotation density, domain-specific resources are less likely to produce irrelevant annotations. On the other hand, domain ontologies are costly to construct as they require substantial manual labor and domain expertise. Consequently, specialized resources are usually less available and more restricted in size: this is why there is currently significant research aiming to enrich ontologies with concept and relation candidates from corpora (Petasis et al., 2011). Therefore, we experimented with combining domain-specific and generic resources to achieve a satisfying balance between annotation density and precision.

3.2. Ontological resources and coverage issues

First, existing ontological/lexical resources were considered for entity annotation. To our knowledge there is no available hand-crafted ontology specific to the NLP domain. Saffron Knowledge Extraction Framework[2] proposes specific terminological resources for different domains, automatically extracted from corpora (Bordea, 2013; Bordea et al., 2013).

Saffron's underlying methodology extracts terms completely automatically in two steps, without relying on comparable corpora. First, high-level terms (*domain models*) are selected and filtered by their weight, with a threshold set empirically for the given corpus. Domain models contain frequent words; they constitute "*the preferred level of naming, that is the taxonomical level at which categories are most cognitively efficient*" (Bordea et al., 2013). We used the available domain models for computer science and for computational linguistics (120 and 200 simple words respectively, with 56 overlapping concepts). Second, intermediate-level terms are extracted according to their distributional similarity to high-level terms. As opposed to the methods based on a comparison between different corpora, this approach favors terms that are less specific but more frequent. The intermediate-level terms (*topic hierarchies*) for computational linguistics, containing 500 units (mostly multioword expressions), were included in our resources. Both high-level and intermediate terms are relatively frequent: Saffron's terms *"are specific to a domain but broad enough to be usable for summarisation or classification"* (Bordea, 2013).

Despite the quality reported for Saffron's resources (Bordea, 2013), the annotation density they provide is definitely too limited for our purposes, especially with respect to pattern mining. An annotation test run on 1.100.000 words yielded an average of 13 annotated concepts/100 words. To increase density, extensive general ontologies were considered with a presumably good coverage for the NLP domain, such as WordNet, BabelNet (Navigli and Ponzetto, 2012) or DBpedia (Mendes et al., 2012). BabelNet being the most extensive (with merged synsets from WordNet, Wikipedia and Wiktionary), it was selected for the annotation experiments. This resource has the advantage of providing significantly increased density - on the other hand, it also increases ambiguity and introduces less relevant annotations. Keeping every BabelNet-annotated entity as a concept candidate was clearly not an option, as test runs showed that more than 80% of words would qualify as a candidate, most of them being irrelevant to the domain. The next section presents the filtering approaches we tried to reduce the number of irrelevant entities.

3.3. Combining Resources with Data-driven Methods

We examined several solutions to filter out relevant annotations from BabelNet while keeping a good annotation density. The first logical step would be to exploit the structure of the resource: in our case, WordNet synsets and Wikipedia categories. By identifying domain-relevant synsets and categories, we could limit the number of entities annotated from BabelNet to those specific to our domain. However, we found that DBpedia classes[3], retrieved from Wikipedia categories are too generic for our purposes. The YAGO ontology (Suchanek et al., 2007) of English Wikipedia classes was considered unfit for our purposes as it is specifically designed for named entities and information extraction. A general problem we encountered about taxonomies of Wikipedia classes stems from the large number of classes and the fact that Wikipedia pages are not linked to classes in a systematic way.

Another solution to limit the number of annotations from BabelNet is to filter concept candidates *before* looking them up in BabelNet. The idea is to combine terminology extraction / multi-word expression (MWE) extraction with ontology lookup. Both vocabulary extraction methods are expected to be useful for filtering: MWEs are presumably more specific than simple words due to the composition and terms are extracted according to domain specificity.

First, we applied the *phrase* tool included in word2vec[4] (Mikolov et al., 2013) to extract multi-word units from the corpus. The abstract part of the corpus was used for training, and was annotated with the identified MWEs. Only the recognised MWEs were looked up in BabelNet. This process yielded a density 3.7 concepts/100 words. This coverage was estimated insufficient.

In the following experiment, the term extraction tool TermSuite (Daille et al., 2013) was applied to the corpus. The extraction process takes as input a set of documents (one abstract/document) and returns a list of term candidates together with a specificity value. Several additional filtering parameters can be applied. We filtered out any candidate whose part of speech is not common noun (or a multi-word unit ending with a common noun). Words that were unknown for the CRF tagger were also excluded, as well as nouns with less than 5 occurrences. Finally, the specificity threshold was set empirically. After setting these parameters, the resulting list was manually validated. The concept candidates coming from terminology extraction were compared with BabelNet concept candidates, and

[2] http://saffron.insight-centre.org/

[3] http://wiki.dbpedia.org/services-resources/ontology

[4] https://code.google.com/p/word2vec/

produced an overlap of 3.805 candidates. The corresponding synsets from BabelNet resources (Wikipedia, WordNet or Wiktionary) were thus selected for entity annotation. This combination resulted in an annotation density of 23 concepts/100 words, which is advantageous for our purposes.

3.4. Evaluation and Error Analysis

We manually evaluated annotation precision on a sample containing 100 sentences, with 358 annotated entities and 932 annotations (an entity is thus linked to 2.6 resources on average). The sample shows the contexts and the annotations together with their sources. Concept candidates were classified as correct or incorrect, with the following error subcategories:

- "non relevant", e.g. :
  written in the early XX century by <entity error="non-relevant">main-stream </entity> authors ...
- "wrong delimitation" : only a part of a multi-word entity is annotated, or each part is annotated as a separate entity, e.g. :
  <entity error="delimitation">Minimum</entity> <entity error="delimitation">Description</entity> <entity error="delimitation">Length</entity>
- "tagging error", e.g. :
  approach that aims to <entity error="tagging"> model</entity>

This categorization of errors takes two different aspects into consideration: the precision of the resource (in terms of relevant/irrelevant annotations produced by the resource), and the precision of the annotation process itself (in terms of candidates that should not be linked to a resource). In other words, we distinguished between error types that could be addressed by improving the corpus pre-processing or the annotation process, and error types coming directly from the resource. Evidently, the distinction is not always clear-cut: semantically ambiguous terms produce both relevant and irrelevant annotations. For instance, the word *term* is often, but not always, used as a domain-relevant concept e.g.:
We show, in <entity> terms </entity> of crossing rates, ...

Beside entity delimitation, the annotation also includes a double reference to the resource: its major type (Saffron/BabelNet) and a minor type (NLP/computer science domain models and topic hierarchies, WordNet, Wikipedia or Wiktionary). This allows us to compute resource-specific precision. Tagging errors were discarded in this evaluation. Table 1 shows the aggregated results for the two major resources in proportion to annotated *expressions*, i.e. if an expression was annotated by more than one minor resource type, they were collapsed. For instance, if the same entity was found in both a Wikipedia Babel synset and a WordNet Babel synset, it counts as a single annotated BabelNet concept. A precision similar to that of Saffron's resources was reached for filtered BabelNet annotations, which confirms the interest of combining data-driven term extraction with large-coverage external resources.

Resource	density /100 w	precision
Saffron - all	13	0.98
BabelNet filtered - all	23	0.97

Table 1: Quantity and precision of entities annotated

Table 2 shows the annotation precision of specific resources (with minor types). Every annotation produced by each specific resource was considered, hence the higher density. POS tagging was still discarded. These data confirm that filtering helped to maintain consistency in the quality of the different annotation resources.

Resource	density /100 w	precision
Saffron concepts	2.7	1
Saffron acl dm	13	0.97
Saffron cs dm	9	0.99
wikipedia	18	0.97
wiktionary	5.5	0.94
wordnet	16	0.98

Table 2: Annotation density and precision for each resource

In terms of annotations produced, the most frequent error type is bad delimitation (62.5% of all errors), followed by POS tagging errors (29.5%), and irrelevant annotations (8%). Finally, Table 3 shows the proportion of different types of errors in terms of annotated expressions. The high proportion of delimitation errors is partly explained by the fact that consecutive parts of a multi-word expression are often annotated as separate instances (see the *Minimum Description Length* example above) and count for as many errors as they have recognized constituents.

Annotation quality	Proportion
Correct	60%
Delimitation error	21%
Tagging error	16%
Irrelevant	3%

Table 3: Error types

4. Unfolding Semantic Relation Types in Scientific Papers

4.1. The typology of Relations

While concepts could be retrieved from existing resources, the relevant semantic relations of the domain were to be identified and annotated manually. The types of relations were defined with a data-driven approach, in parallel to the first round of manual annotation of relation instances. The goal of this step was to study which kinds of relations are present and how they are expressed in scientific papers. Another objective was to verify whether the data confirm the hypothesis behind the pattern mining approach, i.e. that relations between entities are explicit in at least a subset of

Generic Relations	
antonyms	<arg1>similarities</arg1> rather than the <arg2>differences</arg2>
co-hyponyms	<arg1>analysis</arg1> as well as <arg2>synthesis</arg2>
is-a	<arg2>task</arg2> addressed here is that of <arg1>information retrieval</arg1>
Domain-specific relations	
affects	ARG1: *specific property of data* ARG2: *results* <arg1>issues</arg1> that influence the <arg2>effectiveness</arg2>
based_on	ARG1: *method, system* based on ARG2: *other method* <arg1>parser</arg1> based on <arg2>maximum entropy</arg2>
char	ARG1: *observed characteristics* of an ARG2: *entity* <arg1>words</arg1> which occur in the same <arg2>contexts</arg2>
compare	ARG1: *result (of experiment)* to ARG2: *result 2* <arg1>results</arg1> of the experiments are compared with a <arg2>gold standard</arg2>
composed_of	ARG1: *database/resource* ARG2: *data* <arg1>corpus</arg1> consisting of <arg2>sentences</arg2>
datasource	ARG1: *information* extracted from ARG2: *data* <arg1>word</arg1> in both languages is extracted from the <arg2>corpus</arg2>
methodapplied	ARG1: *method* applied to ARG2: *data* <arg1>approach</arg1> is illustrated by applying it to large <arg2>corpora</arg2>
model	ARG1: *abstract representation* of an ARG2: *observed entity* <arg1>parse tree</arg1> representation of the input <arg2>sentences</arg2>
phenomenon	ARG1: *phenomenon* found in ARG2: *context* <arg1>differences</arg1> attested among <arg2>languages</arg2>
problem	ARG1: *phenomenon* is a problem in a ARG2: *field* <arg1>ambiguity</arg1> exists in the <arg2>input</arg2>
propose	ARG1: *paper/author* presents ARG2: *an idea* <arg1>paper</arg1> describes a <arg2>framework</arg2>
study	ARG1: *analysis* of a ARG2: *phenomenon* <arg1>study</arg1> of <arg2>word meaning</arg2>
tag	ARG1: *meta-information* associated to ARG2: *entity* <arg1>corpus</arg1> annotated for <arg2>semantic information</arg2>
taskapplied	ARG1: *task* performed on ARG2: *data* <arg1>tagging</arg1> English <arg2>texts</arg2>
usedfor	ARG1: *method/system* ARG2: *task* <arg1>method</arg1> we used to <arg2>search</arg2>
uses_information	ARG1: *method* relies on ARG2: *information* <arg1>technique</arg1> relies on explicit <arg2>relevance</arg2>
yields	ARG1: *experiment/method* ARG2: *result* <arg1>method</arg1> achieved better <arg2>accuracy</arg2>
wrt	ARG1 *a change* in/with respect to ARG2: *property* <arg1>improvement</arg1> in <arg2>translation quality</arg2>

Table 4: Semantic Relation Typology

entity mention pairs. First, semantic relation types were identified in a sample of 100 abstracts extracted from the corpus.

Relation	Frequency in corpus
usedfor	27%
composed_of	16%
propose	11%
yields	6%
study	6%
taskapplied	5%
uses_information	4%
affects	4%

Table 5: Most frequent semantic relations

On the textual level, a semantic relation will be conceived as a text span linking two annotated instances of concepts within the same sentence. Only explicit relations were taken into account, i.e. examples when it is realistic to expect a sequential data mining algorithm to recognize and annotate the sequence as an instance of a relation. If the two entities are in a semantic relation with each other but this relation is not expressed in the text, the instance was not annotated.

The annotation covers the text span between the two entities, and specifies the type of the relation, as well as the two arguments. Sequences can contain gaps: not every word in the context is expected to be relevant for the relation. In the example below, only the highlighted text span

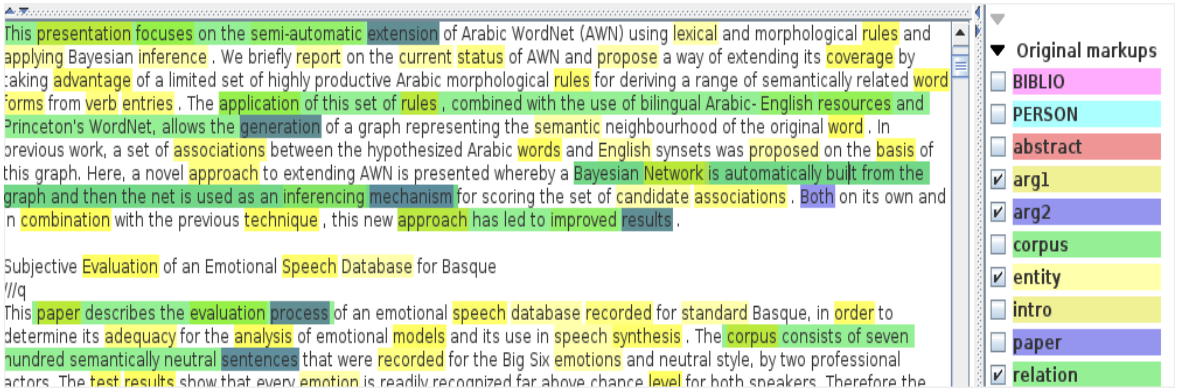

Figure 1: Manual annotation with the GATE interface

is actually informative for identifying the semantic relation:

<relation type="usedfor"><arg1><entity> scenarios </entity></arg1> can be simulated at a preliminary stage, instead of real-time implementations, **allowing for** repeatable <arg2><entity>experiments</entity> </arg2></relation>.

On the semantic level, a relation type needs to be specific enough to be easily defined and identified for a domain expert. As it was revealed by the manual annotation, in order to achieve this level of specificity, the arguments of a relation have to be typed, too. If a concept is linked to a WordNet synset from BabelNet, this type can eventually correspond to the synset description. Table 4.1. presents the typology of relations defined at the first step, together with argument type specifications [5].

4.2. Annotation of Relations

After the typology of relations was defined, 87 instances of relations were manually annotated with GATE (Cunningham et al., 2002) in a sample of 100 abstracts. A larger sample of 500 abstracts, following the distribution of the corpus with respect to the source of the paper (conference/workshop type), is to be annotated by two independent annotators and will serve as training/evaluation corpus for relation extraction experiments.

One of the major issues in associating semantic relations to an explicit text span is the problem of overlapping relations, including those with more than two arguments. A single concept can be the argument of more than one (explicit) relations in the same context. For instance, consider the following sentence:

This paper presents an application of Text Zoning to the ACL Anthology.

Two instances of relations can be recognized according to our semantic relation typology (Figure 4.1.):

propose (ARG1: paper ARG2: TextZoning)
taskapplied(ARG1: Text Zoning ARG2: ACL Anthology)

[5]The final typology may be subject to modifications as long as the manual annotation phase is not finished

However, the two relation instances are overlapping:

- This <relation type="propose">paper presents an application of Text Zoning</relation>to the ACL Anthology.
- This paper presents an <relation type="taskapplied">application of Text Zoning to the ACL Anthology</relation>.

In these cases, the most relevant binary relation has to be selected for annotation.

The findings of this work phase are twofold. First, as shown by the examples above, the semantic relations are not specific to natural language processing but reflect the more general semantics of the science/engineering domain, although particular arguments of the relations in the corpus can be more domain-specific. Second, entities participating in relations are mostly high level concepts. E.g., typical instances of the relation *affects* include the following arguments: *properties* affect *results*; *question* affects *performance*, *type* affects *differences*. One reason behind this may be that abstracts and introductions aim to put the content in perspective instead of presenting precise details. Another explanation can be that our concept resources prioritize high to intermediate level terminology (Bordea et al., 2013) as the "preferred level of naming". This feature is advantageous for pattern mining.

Manual annotation also revealed some limitations due to conscious choices. An issue we had foreseen concerns anaphoric expressions. Although we consciously exclude relations between entities expressed as pronouns, this does not seem to result in significant loss of information, as abstracts and introductions occur early in the paper and usually contain the first mention of an entity.

However, other limitations stem from entity annotation errors, in particular bad delimitation. These errors affect the annotation and especially the quality of the relation instances to be retrieved automatically.

5. Conclusion and Future Work

We have presented the first experiments and outcome of annotating the ACL Anthology with domain-relevant concepts and semantic relations. Our studies on concept

annotation confirmed that ontology specificity correlates with annotation precision but comes with limited coverage. The experiments confirmed that data-driven term extraction methods can be efficiently combined with large-coverage external resources, and allow to expand coverage while maintaining the same level of precision that is achieved using domain-specific resources. As a result, the entire corpus was automatically annotated for domain concepts.

We also presented a typology of semantic relations in the science/engineering domain and are carrying out a manual annotation experiment. The first findings on applying the typology to the corpus are encouraging in that valuable information can be identified and the limitations revealed by the error analysis can further be addressed.

This corpus is currently being exploited for experiments in unsupervised relation extraction. Sequence mining methods are used to identify relevant text patterns between concepts. The distribution of concept couples over sequential patterns allow to cluster the instances according to their semantic relation. We also expect to be able to discover new, domain-specific relation types via unsupervised clustering. The next scheduled step is the manual annotation of semantic relations in 500 abstracts. This work should allow to gain information on the plausibility and usability of our relation typology by measuring inter-annotator agreement. The resulting annotated corpus will be shared with the research community, allowing to compare relation extraction algorithms.

6. Acknowledgments

This work is part of the program "Investissements d'Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL).

The authors are grateful to Elisa Omodei for sharing the preprocessed corpus of abstracts from the ACL Anthology, and to Georgeta Bordea for putting the ACL domain models at our disposal.

7. Bibliographical References

Béchet, N., Cellier, P., Charnois, T., and Crémilleux, B. (2012). Discovering linguistic patterns using sequence mining. In *Computational Linguistics and Intelligent Text Processing - 13th International Conference, CICLing*, pages 154–165.

Bordea, G., Buitelaar, P., and Polajnar, T. (2013). Domain-independent term extraction through domain modelling. In *10th International Conference on Terminology and Artificial Intelligence (TIA 2013)*.

Bordea, G. (2013). *Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining*. Phd thesis, National University of Ireland, Galway.

Brewster, C., Alani, H., and Dasmahapatra, A. (2004). Data driven ontology evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Chavalarias, D. and Cointet, J.-P. (2013). Phylomemetic patterns in science evolution - the rise and fall of scientific fields. *PLOS ONE*, 8(2).

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the ACL 2002 Conference*.

Daille, B., Jacquin, C., Monceaux, L., Morin, E., and Rocheteau, J. (2013). TTC TermSuite : Une chaîne de traitement pour la fouille terminologique multilingue. In *Proceedings of the Traitement Automatique des Langues Naturelles Conference (TALN)*.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge (MA).

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the ACL Conference*.

Herrera, M., Roberts, D. C., and Gulbahce, N. (2010). Mapping the evolution of scientific fields. *PLOS ONE*, 5.

Larsen, P. O. and von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index.

Mendes, P., Jakob, M., and Bizer, C. (2012). Dbpedia for nlp: A multilingual cross-domain knowledge base. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR, 2013.*

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Omodei, E., Cointet, J.-P., and Poibeau, T. (2014). Mapping the natural language processing domain : Experiments using the acl anthology. In *International Conference on Language Resources and Evaluation LREC*.

Osborne, F. and Motta, E. (2012). Mining semantic relations between reserach areas. In *International Semantic Web Conference, Boston (MA)*.

Osborne, F. and Motta, E. (2014). Rexplore: unveiling the dynamics of scholarly data. *Digital Libraries*, 8(12).

Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., and Zavitsanos, E. (2011). Ontology population and enrichment: State of the art. In *Knowledge-driven multimedia information extraction and ontology evolution*, pages 134–166. Springer-Verlag.

Presutti, V., Consoli, S., Nuzzolese, A. G., Recupero, D. R., Gangemi, A., Bannour, I., and Zargayouna, H. (2014). Uncovering the semantics of wikipedia pagelinks. In *Knowledge Engineering and Knowledge Management*, pages 413–428. Springer.

Radev, D., Muthukrishnan, P., and Qazvinian, V. (2009). The acl anthology network corpus. In *Proceedings of the 2009 ACL Workshop on Text and Citation Analysis for Scholarly Digital Libraries*.

Sagot, B. and Boullier, P. (2008). Sxpipe 2: architecture pour le traitement pré-syntaxique de corpus bruts. *Traitement Automatique des Langues*, 49 (2).

Sateli, B. and Witte, R. (2015). What's in this paper?: Combining rhetorical entities with linked open data for

semantic literature querying. In *Proceedings of the 24th International Conference on World Wide Web*.

Shotton, D. (2010). CiTO, the Citation Typing Ontology. *J. Biomedical Semantics*, 1(S-1).

Skupin, A. (2004). The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences*, 101.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago - a core of semantic knowledge. In *Proceedings of the 16th International World Wide Web Conference*.

Wang, W., Besançon, R., Ferret, O., and Grau, B. (2013). Extraction et regroupement de relations entre entités pour l'extraction d'information non supervisée. *Traitement automatique de la langue*, 54.